Regular languages and finite automata.

Suppose Σ is a finite alphabet.

Recall the following basic operations on languages $L, K \subseteq \Sigma^*$:

Concatenation

$$LK = \{xy \mid x \in L, y \in K\}$$

Union

$$L \cup K = \{x \mid x \in L \text{ or } x \in K\}$$

Kleene star

$$L^* = L^0 \cup L^1 \cup L^2 \cup \dots$$
$$= \{ w_1 w_2 \dots w_n \mid n \geq 0 \ \& \ w_i \in L \}.$$

Definition The class of regular languages in Σ is the smallest class $\mathcal{R}$ of languages containing $\emptyset$, $\{a\}$ (for $a \in \Sigma$) and such that if $L, K \in \mathcal{R}$, then also $LK, L \cup K, L^* \in \mathcal{R}$.

Example All finite languages are regular, since
$\{a_1 a_2 \dots a_n\} = \{a_1\} \circ \{a_2\} \circ \dots \circ \{a_n\}$.

<u>Definition</u> A <u>deterministic</u> <u>finite</u> <u>state</u> <u>automaton</u>, DFA, consists of a finite directed graph $\mathcal{M}$, i.e., $\mathcal{M} = (V, E)$ where V is a finite set of <u>vertices</u> and $E \subseteq V^2$ is a set of directed <u>edges</u>, along with

(1) a distinguished <u>start</u> <u>state</u> $s_0 \in V$,

(2) a set $A \subseteq V$ of <u>accepting</u> <u>states</u>,

(3) a labelling $\ell : E \to \mathcal{P}(\Sigma) \setminus \{\emptyset\}$

for any $s \in V$ and any $a \in \Sigma$ there is exactly one edge $(s, t) \in E$ originating at s and with label a, i.e., $a \in \ell(s, t)$.

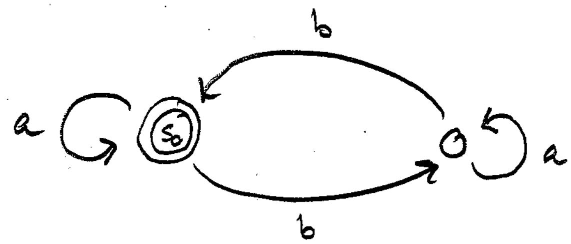
We call V the set of <u>states</u> of the automaton.

Given an automaton $\mathcal{M}$ as above and a string $w = a_1 a_2 \ldots a_n \in \Sigma^*$, we say that $\mathcal{M}$ <u>accepts</u> w if the unique edgepath $(e_1, e_2, \ldots, e_n)$ in $\mathcal{M}$ originating at s_0 and with $a_j \in \ell(e_j)$ terminates at an accepting state.

Given $\mathcal{M}$, we let $L(\mathcal{M})$ be the language consisting of the strings accepted by $\mathcal{M}$. We also say that $\mathcal{M}$ <u>recognises</u> $L(\mathcal{M})$.

<u>Example</u> We draw a DFA by a diagram in the plane as a usual directed graph with labeled edges. Moreover, the accepting states are indicated by double circles.

$\mathcal{M}$:

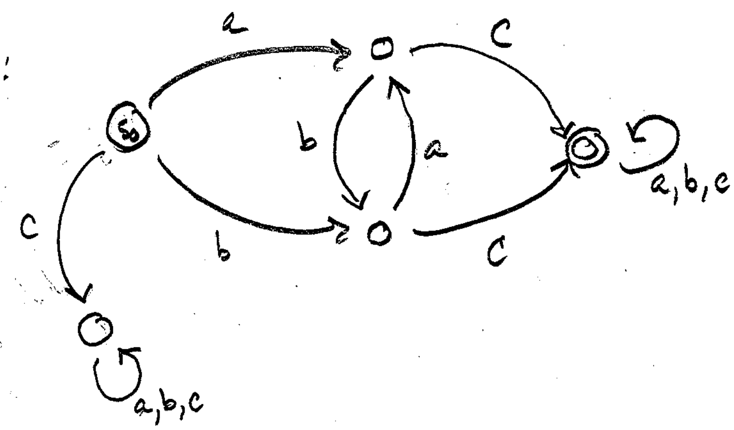


Here $\mathcal{M}$ has exactly two states and the only accepting state is s_0. We see that $\Sigma = \{a, b\}$ and

$$L(\mathcal{M}) = \{ w \mid |w|_b = \# \text{ of occurrences of } b \text{ in } w \text{ is even} \}.$$

Similarly, for $\Sigma = \{a, b, c\}$

$\mathcal{N}$:



$$L(\mathcal{N}) = \{ awcv, bwcv \mid w \in \{a,b\}^*, v \in \{a,b,c\}^* \}.$$

Definition A *generalised non-deterministic finite state automaton*, ε-NFA, is a directed finite graph $\mathcal{M} = (V, E)$ along with

(1) a non-empty set of start states $S \subseteq V$,

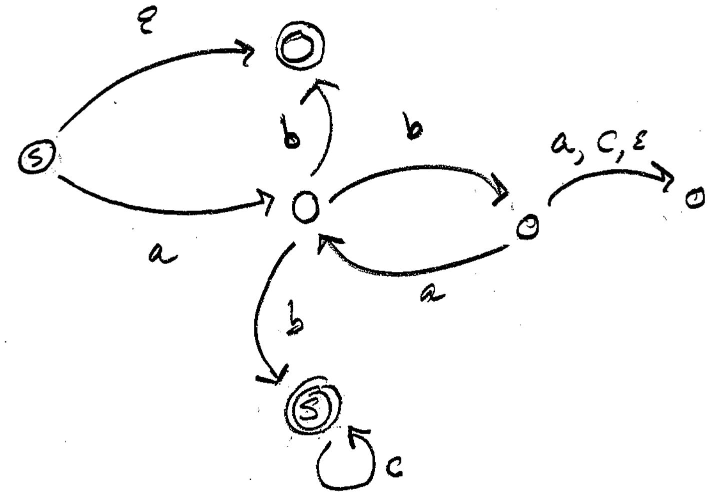
(2) a set of accepting states $A \subseteq V$,

(3) a labelling $\ell : E \to \mathcal{P}(\Sigma \cup \{\varepsilon\}) \setminus \{\emptyset\}$.

Given an ε-NFA $\mathcal{M}$ and a string $w \in \Sigma$, we say that $\mathcal{M}$ accepts w if there is an edge path $(e_1, \dots, e_m)$ and labels $b_i \in \ell(e_i)$ (b_i can be ε) such that $w = b_1 b_2 \cdots b_m$.

Example We indicate the start states by an S.

$\mathcal{M}$



$L(\mathcal{M}) = \{ (ab)^n c^m \mid n \geq 0, m \geq 0 \}$

Cone types

Suppose $L \subseteq \Sigma^*$ is a language and $w \in \Sigma^*$.

We define the cone type of w wrt L by

$$\mathrm{Cone}_L(w) = \{ x \in \Sigma^* \mid wx \in L \}$$

and let the cone types of L be

$$\mathrm{Cone}(L) = \{ \mathrm{cone}_L(w) \mid w \in \Sigma^* \}$$

Note The set of cone types of L are the cones of all words $w \in \Sigma^*$, not just the $w \in L$.

Example Let $L = \{a^n b^n \mid n \geq 0\} \subseteq \{a,b\}^*$.

Then $\mathrm{cone}_L(b^2) = \emptyset$, $\mathrm{cone}_L(\varepsilon) = L$,

$\mathrm{cone}_L(a^2) = \{ a^n b^{n+2} \mid n \geq 0 \}$.

Note also that $\mathrm{cone}_L(a)$, $\mathrm{cone}_L(a^2)$, ---- are all distinct.

Definition For $L \subseteq \Sigma^*$ a language define an equivalence relation $\sim_L$ on Σ^* by

$w \sim_L v \iff w$ and v have the same cone type

i.e., $\mathrm{cone}_L(w) = \mathrm{cone}_L(v)$.

Lemma Let $\mathcal{M}$ be an ε-NFA and let $L = L(\mathcal{M})$ be the language accepted by $\mathcal{M}$. Then L has only finitely many cone-types, i.e., $\sim_L$ has only finitely many classes.

Proof We define another equivalence relation $\approx$ on Σ^* as follows:

$w \approx v \iff$ for any start state s and arbitrary state t of $\mathcal{M}$, there is an edge path from s to t with edge label w if and only if there is an edge path from s to t with edge label v.

Note then ~~that~~ if $\mathcal{M}$ has n states, then $\approx$ has at most 2^{n^2} classes (for each edge (s,t), we have to respond to a yes-no question).

Claim If $w \approx v$ then also $w \sim_L v$.

For suppose that, e.g., $x \in \mathrm{cone}_L(w)$. Then there is an edge path beginning at a start state s and terminating at an accepting state t_0 with edge label wx. Let t_1 be any state along this edge path such

$$\underbrace{\overset{\substack{\text{beginning at } s\\ \downarrow}}{e_1} e_2 \cdots\cdots \overset{\substack{\text{state } t_1\\ \downarrow}}{e_m}}_{\text{label } w} \underbrace{e_{m+1} \cdots\cdots \overset{\substack{\text{terminating at } t_0\\ \downarrow}}{e_k}}_{\text{label } x}$$

that the edge path arrives at t_1 with label w. Then there is an alternative edge path $e_1' \cdots e_p'$ beginning at s and terminating at t_2 with label v. It follows that there is an edge path from s to t_0 with label vx, whence $x \in \text{cone}_L(v)$.

$$\underbrace{\overset{\substack{\text{beginning at } s\\ \downarrow}}{e_1'} \cdots \overset{\substack{\text{state } t_1\\ \downarrow}}{e_p'}}_{\text{label } v} \underbrace{e_{m+1} \cdots \overset{\substack{\text{terminating at } t_0\\ \downarrow}}{e_k}}_{\text{label } x}$$

Similarly, $\text{cone}_L(v) \subseteq \text{cone}_L(w)$. $\square$

<u>Lemma</u> Suppose $L \subseteq \Sigma^*$ is a language with finitely many cone types. Then there is a DFA $\mathcal{M}$ with $L(\mathcal{M}) = L$.

<u>Proof</u> Let $\Delta_1, \ldots, \Delta_n$ be the finitely many cone types of L and note that for $w, x \in \Sigma^*$ and $a \in \Sigma$, if $w \sim_L x$ then also $wa \sim_L xa$.

Let $\mathcal{M}$ have states $\Delta_1, \ldots, \Delta_n$ and put an arrow $\Delta_i \xrightarrow{a} \Delta_j$ if for some/

any $w \in \Sigma^*$ with $\mathrm{cone}_L(w) = \Delta_i$, we have $\mathrm{cone}_L(wa) = \Delta_j$. Then a state Δ_i is accepting if $\varepsilon \in \Delta_i$ and $s = \mathrm{cone}_L(\varepsilon)$ is the unique start state. $\mathcal{M}$ is clearly deterministic and $L(\mathcal{M}) = L$. $\square$

<u>Lemma</u> Let $\mathcal{M}$ be a DFA. Then the language $L(\mathcal{M})$ is regular.

<u>Proof</u> Let V be the finite set of states of $\mathcal{M}$ and let $t_0, t_1 \in V$ be arbitrary. For any $X \subseteq V$, let

$G(X, t_0, t_1) = \{ w \in \Sigma^* \mid$ there is an edge path from t_0 to t_1 only passing through states in X and having label $w \}$.

By induction on $|X|$, we show that for any t_0, t_1, the language $G(X, t_0, t_1)$ is regular.

$|X| = 0$: In this case $X = \emptyset$ and so $w \in G(X, t_0, t_1)$ if and only if $w = a$ for some $a \in \Sigma$ for which there is an edge $t_0 \xrightarrow{a} t_1$.

So $G(X, t_0, t_1)$ is a (finite) subset of Σ

and hence is regular.

$|X| = n+1$: Assume the result holds for all subsets of V of size $\leq n$ and assume $|X| = n+1$.

Then we have that

$$G(X, t_0, t_1) = \left(\bigcup_{q \in X} G(X \setminus \{q\}, t_0, t_1) \right)$$

$$\cup \left(\bigcup_{q \in X} G(X \setminus \{q\}, t_0, q) \circ G(X \setminus \{q\}, q, q)^* \circ G(X \setminus \{q\}, q, t_1) \right)$$

which is regular by the induction hypothesis.

Clearly, the right hand side is contained in $G(X, t_0, t_1)$.

Conversely, suppose $w \in G(X, t_0, t_1)$ and consider the edge path from t_0 to t_1 with label $w = a_1 a_2 \cdots a_n$:

$$t_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} q_2 \longrightarrow \cdots \xrightarrow{a_{n-1}} q_{n-1} \xrightarrow{a_n} t_1 .$$

Then, if $|w| = n \leq 1$, we have $w \in \bigcup_{q \in X} G(X \setminus \{q\}, t_0, t_1)$.

Otherwise, note that

- $a_1 \in G(X \setminus \{q_1\}, t_0, q_1)$
- if q_i is the last occurrence of q_1 among $q_1, \ldots, q_{n-1}$, then

$a_{i+1} \cdots a_n \in G(X \setminus \{q_1\}, q_1, t_1)$

- if $q_i = q_j = q_1$ for $i < j$ and $q_\ell \neq q_1$ for all $i < \ell < j$, then

$$a_i a_{i+1} \cdots a_{j-1} \in G(X \setminus \{q_1\}, q_1, q_1).$$

Thus, $w = a_1 \cdots a_n$ belongs to

$$G(X \setminus \{q_1\}, t_0, q_1) \circ G(X \setminus \{q_1\}, q_1, q_1)^* \circ G(X \setminus \{q_1\}, q_1, t_1). \blacktriangle$$

So also $L(\mathcal{M}) = \bigcup_{q \in A} G(V, s_0, q)$ is regular. $\square$

Theorem TFAE for a language L over a finite alphabet Σ.

(a) L is regular.

(b) $L = L(\mathcal{M})$ for some DFA $\mathcal{M}$.

(c) $L = L(\mathcal{M})$ for some ε-NFA $\mathcal{M}$

(d) $\mathrm{Cone}(L)$ is finite.

Here the equivalence of regular languages and languages recognised by finite automata is known as the Kleene theorem, while the equivalence with (d) is the Myhill-Nerode theorem.

<u>Proof</u> We have already proved (c) $\Rightarrow$ (d) $\Rightarrow$ (b) $\Rightarrow$ (a).

So we need only prove that regular languages are of the form $L(\mathcal{M})$ for ε-NFA $\mathcal{M}$.

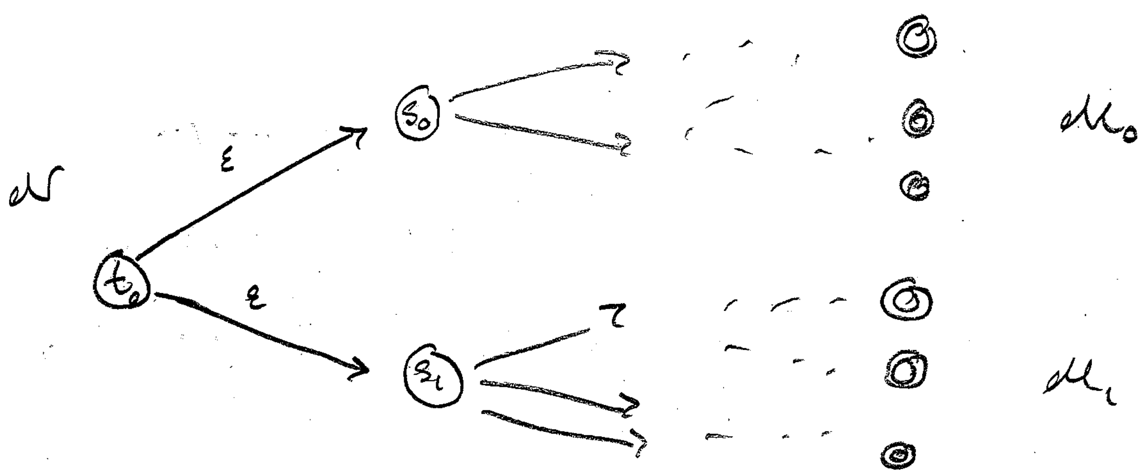
Since one can easily build ε-NFA recognising any finite language, it suffices to show that if L, K are recognised by DFA, then also LK, $L \cup K$ and L^* are recognised by ε-NFA.

So suppose $\mathcal{M}_0 = (V_0, E_0, s_0, A_0, \ell_0)$ and $\mathcal{M}_1 = (V_1, E_1, s_1, A_1, \ell_1)$ are deterministic finite automata recognising L and K respectively. That is, V_i are the states, $E_i \subseteq V_i^2$ the directed edges, s_i the start state, A_i the accepting states and $\ell_i : E_i \longrightarrow \mathcal{P}(\Sigma) \setminus \{\emptyset\}$ the labelling. <u>Wlog</u>, $V_0 \cap V_1 = \emptyset$.

Let us first build an ε-NFA recognising $L \cup K$:

Set $\mathcal{N} = \big(V_0 \cup V_1 \cup \{t_0\},\ E_0 \cup E_1 \cup \{(t_0, s_0), (t_0, s_1)\},\ t_0,\ A_0 \cup A_1,\ \ell \big)$,

where $\ell(e) = \ell_0(e)$ for $e \in E_0$, $\ell(e) = \ell_1(e)$ for $e \in E_1$ and $\ell(t_0, s_0) = \ell(t_0, s_1) = \varepsilon$.



Thus, at the first stage of the computation, w/o reading any input of the string, $\mathcal{N}$ has to decide to feed the input to either $\mathcal{M}_0$ or $\mathcal{M}_1$. So $L(\mathcal{N}) = L \cup K$.

Now, let us construct $\mathcal{N}$ to recognize LK.

$$\mathcal{N} = (V_0 \cup V_1,\ E_0 \cup E_1 \cup \{(q, s_1) \mid q \in A_0\},\ s_0,\ A_1,\ \ell)$$

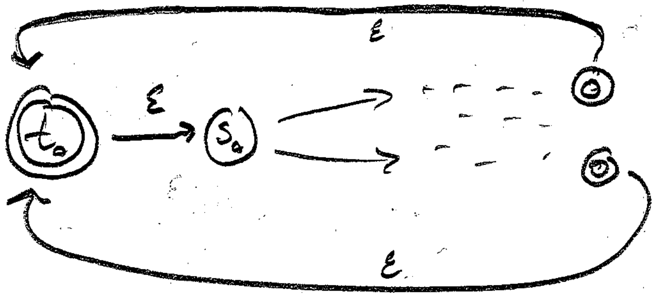
where $\ell(e) = \ell_i(e)$ for $e \in E_i$ and $\ell(q, s_1) = \varepsilon$ for $q \in A_0$. Then, on input w, $\mathcal{N}$ begins with a computation in $\mathcal{M}_0$ and can from any accepting state of $\mathcal{M}_0$ jump to s_1 and then continue with a computation in $\mathcal{M}_1$. So $\mathcal{N}$ accepts w if and only if $w = xy$, where $x \in L$ and $y \in K$. Thus, $L(\mathcal{N}) = LK$.

Finally, for L^* let t_0 be a new state and set

$$\mathcal{N} = \Big(V_0 \cup \{t_0\}, E_0 \cup \{(t_0, s_0)\} \cup \{(q, t_0) \mid q \in A_0\}, t_0, A_0 \cup \{t_0\}, \ell \Big),$$

where $\ell(e) = \ell_0(e)$ for $e \in E_0$, $\ell(t_0, s_0) = \varepsilon$, $\ell(q, t_0) = \varepsilon$ for $q \in A_0$.

Again, $L(\mathcal{N}) = L^*$.



$\square$

<u>Remark</u> We can of course also show that regular languages are recognised by DFA by instead using Myhill-Nerode theorem, i.e., by showing that they only have finitely many cone types.

For example, suppose L and K have finitely many cone types each and let $\Delta_1, \ldots, \Delta_n$ be the cone types of K.

Now set

$$w \approx v \iff w \sim_L v$$

& for any $i \le n$, w has a decomposition as $w = xy$ with $x \in L$ and $\mathrm{cone}_K(y) = \Delta_i$ if and only if v has.

Clearly, $\approx$ is an equivalence relation with finitely many classes (note that the second part is induced by n yes/no questions). Moreover, suppose that $w \approx v$ and that $wu \in LK$. Then either we can write $u = xy$, where $y \in K$ and $wx \in L$, whence also $vx \in L$ and thus $vu = vxy \in LK$, or we can write $w = xy$, where $x \in L$ and $yu \in K$. In the second case, let $\Delta_i = \mathrm{cone}_K(y)$ and note that then v has a decomposition as $v = st$, where $s \in L$ and $\mathrm{cone}_K(t) = \Delta_i$. It thus follows that $tu \in K$, whence $vu = stu \in LK$.

In any case, $vu \in LK$, so $\mathrm{cone}_{LK}(w) \subseteq \mathrm{cone}_{LK}(v)$. By symmetry, $w \sim_{LK} v$. So $\approx$ refines $\sim_{LK}$ and hence LK has finitely many cone types.

<u>Exercise</u>: Show that if L, K have finitely many cone types, then also $L \cup K$ and L^* have finitely many cone types.

Before giving more examples of regular languages, we give some indication of their limitations.

Pumping lemma

Let $L \subseteq \Sigma^*$ be a regular language. Then there is an integer $n \geq 1$ such that any word $w \in L$ with $|w| > n$ can be expressed as

$$w = xyz \quad , \; y \in \Sigma^+ \; , \; x, z \in \Sigma^*$$

where $|x| \leq n$ and $xy^iz \in L$ for all $i \geq 0$.

Proof

First suppose L is recognised by a DFA $\mathcal{M}$ with n states. Then if w is a word of length $> n$ the edge path labeled by w will have to pass through some state twice. So let x be the shortest prefix of w at which a state q is reached that is being visited twice. Clearly $|x| \leq n$. Also, let y be the shortest non-empty string s.t. the edge path with label xy arrives to q. Then so does xy^i for all $i \geq 0$ and thus $xy^iz \in L$ for all $i \geq 0$. $\square$

Example The languages $\{a^n b^n \mid n \geq 0\}$ and $\{a^n \mid n \text{ is a prime}\}$ are not regular.

The subsequence ordering

<u>Definition</u> Let $<$ be a strict partial ordering on a set X, i.e., $<$ is transitive and irreflexive. We say that $<$ is a <u>well-quasi-ordering</u>, wqo, if

(i) any antichain is finite, i.e., for any infinite sequence $(x_i)_{i\in\mathbb{N}}$ there are $i \neq j$ s.t. $x_i = x_j$ or $x_i < x_j$,

(ii) there is no infinite descending chain, i.e., no infinite sequence $(x_i)_{i\in\mathbb{N}}$ with $x_1 > x_2 > x_3 > \cdots$ ($<$ is <u>well-founded</u>)

<u>Exercise</u> Set $x \leq y \iff x < y$ or $x = y$.
Show that $<$ is wqo if and only if for any infinite sequence $(x_i)_{i\in\mathbb{N}}$ there are $i < j$ with $x_i \leq x_j$.
[Hint: Use Ramsey's theorem].

<u>Definition</u> Suppose $w = a_1 a_2 \cdots a_n$ and x are words in Σ. We say that w is a <u>subsequence</u> of x if there are words $y_0, \ldots, y_n \in \Sigma^*$ with $x = y_0 a_1 y_1 a_2 y_2 \cdots a_n y_n$ and $w \neq x$. Write $w \prec x$ to denote this.

Proposition Let Σ be a finite alphabet. Then $(\Sigma^*, <)$ is a well-quasi-ordering.

Proof Assume towards a contradiction that there is some sequence $(x_i)_{i\in\mathbb{N}}$ s.t. $i<j \Rightarrow x_i \not< x_j$ and call any such sequence **bad**.

We inductively construct a bad sequence as follows:

- Let $y_1 \in \Sigma^*$ be a word of shortest length beginning an infinite bad sequence.
- Let $y_2 \in \Sigma^*$ be a word of shortest length s.t. y_1, y_2 begins an infinite bad sequence.
- Let $y_3 \in \Sigma^*$ be a word of shortest length beginning an infinite bad sequence, etc.

Then, as Σ is finite, there is an infinite subsequence, say $(y_{n_i})_{i\in\mathbb{N}}$, with constant first letter, e.g., $y_{n_i} = a z_{n_i}$. Note again that $z_{n_i} \not< z_{n_j}$ for $i<j$ and $z_{n_i} < y_{n_i}$, so

$$y_1, y_2, \ldots, y_{n_1-1}, z_{n_1}, z_{n_2}, \ldots$$

is also an infinite bad sequence, but with $|z_{n_1}| < |y_{n_1}|$, contradicting the minimality of

$$y_1, y_2, \ldots, y_{n_1-1}, y_{n_1}, y_{n_1+1}, \ldots \qquad \square$$

<u>Definition</u> Let $(X, <)$ be a strict partial ordering and $B \subseteq Y \subseteq X$ subsets. We say that B is a <u>basis</u> for Y if

$$Y = \{ x \in X \mid \exists z \in B \ \ z \leq x \}.$$

<u>Exercise</u> Show that $(X, <)$ is wqo if and only if any $Y \subseteq X$, which is closed upwards, i.e., $(y \leq x \ \& \ y \in Y) \Longrightarrow x \in Y$, has a finite basis.

<u>Corollary</u> Let $L \subseteq \Sigma^*$ be a language closed under supersequences, i.e., if $x \in L$ and $x \preceq y$, then $y \in L$.

Then L has a finite basis $B = \{x_1, \dots, x_n\}$, i.e., $L = \{ y \in \Sigma^* \mid x_i \preceq y \text{ for some } i = 1, \dots, n \}$.

<u>Theorem</u> Let $L \subseteq \Sigma^*$ be any language closed under supersequences. Then L is regular.

<u>Proof</u> Let $B = \{x_1, \dots, x_n\}$ be a finite basis for L and let C be the finite set of all prefixes of elements in B.

Set $w \approx v \iff$ for any $x \in C$, $x \preceq w$ if and only if $x \preceq v$.

Then $w \approx v \Rightarrow w \sim_L v$, and since $\approx$ has only finitely many classes, so does $\sim_L$, whence L is regular. $\square$

<u>Theorem</u> Let $L \subseteq \Sigma^*$ be regular, then so is $\Sigma^* \setminus L$.

<u>Pf</u> Note ~~that~~ $\mathrm{cone}_{\Sigma^* \setminus L}(w) = \Sigma^* \setminus \mathrm{cone}_L(w)$, so if L has only finitely many cone types, the same holds for $\Sigma^* \setminus L$. $\square$

<u>Corollary</u> Let $L \subseteq \Sigma^*$ be a language closed under taking subsequences. Then L is regular.

<u>Pf</u> Just note that $\Sigma^* \setminus L$ is closed under supersequences, so $\Sigma^* \setminus L$ and thus also L are regular. $\square$

<u>Example</u> Let $\Sigma = \{0, 1, \dots, 9\}$ and let $L \subseteq \Sigma^+$ be the set of all prime numbers written in base 10 and $K = \{x \in \Sigma^+ \mid \exists y \in L \; y \preccurlyeq x\}$. Then K has a finite basis $B = \{x_1, \dots, x_n\}$, where $x_1, \dots, x_n$ are actually prime numbers written in base 10.

Example Let $\Sigma = \{0,1,2\}$ and let $L \subseteq \Sigma^+$ be the set of all prime numbers written in base 3. Then $\{2, 10, 111\}$ is a basis for $\text{sup}(L) = \{y \in \Sigma^+ \mid \exists x \in L \ \ x \preccurlyeq y\}$.

To see this, note first that $2 \sim 2$, $10 \sim 3$, $111 \sim 13$ (where $2, 3, 13$ are in base 10). So $2, 10, 111 \in L$.

Also, suppose $x \in L$. Then if $2 \not\preccurlyeq x$, we have $x \in \{0,1\}^*$, and if further $10 \not\preccurlyeq x$, also $x \in 0^* 1^*$. Now, unless x represents the number 0, x has no leading 0 and hence $x \in 1^*$. So suppose $x \in 1^*$. Then as neither 1, nor 11 represent prime numbers, we must have $x \in 1111^*$, i.e., $111 \preccurlyeq x$.

Notation Let $x \sqsubseteq y$ if x is a prefix of y.

Theorem Suppose L is a regular language. Then $\text{pref}(L) = \{x \in \Sigma^* \mid \exists y \in L \ \ x \sqsubseteq y\}$ is regular too.

Proof Exercise. $\square$